Assessor Training: Influence of Training Strategy and Perceived Purpose of the Assessment on Overall Rating Accuracy

CHEUNG, Wing Ying

A thesis submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Philosophy

in

Industrial-Organizational Psychology

The Chinese University of Hong Kong

July 2011


香港中文大學
大學
圖書館系統
18 SEP 2013
UNIVERSITY
LIBRARY SYSTEM
THE CHINESE UNIVERSITY OF HONG KONG

Thesis/Assessment Committee

Professor Darius Kwan-shing Chan (Chair)
Professor Winton Wing-tung Au (Thesis Supervisor)
Professor Timothy Carey (Thesis Co-supervisor)
Professor Wai Chan (Committee Member)
Professor Norbert Tanzer (External Examiner)

Abstract of thesis entitled:

Assessor Training: Influence of Training Strategy and Perceived Purpose of the Assessment on Overall Rating Accuracy

Submitted by CHEUNG, Wing Ying

For the Degree of Master of Philosophy in Industrial-Organizational Psychology

at The Chinese University of Hong Kong in July 2011

The present study examined the effects of assessor training strategy and the perceived purpose of the assessment on overall rating accuracy (i.e., inter-rater reliability, correlation accuracy and deviation accuracy). Assessor training strategy was manipulated with four conditions: behavioral observation training strategy (BOT), frame-of-reference training strategy (FOR), combination training strategy (BOT & FOR) and no-training condition. The perceived purpose of the assessment was manipulated in three conditions: personnel selection, developmental feedback and a research-only condition. Two hundred forty university students in Hong Kong were randomly assigned to one of the experimental conditions, and rated two hypothetical videotaped interviews. The expert panel consisted of four experienced human resources practitioners, whose ratings were used as the "standard scores" to compare with the participants' ratings and generate the participants' accuracy indices.

Only the assessor training strategy was found to be significant factor; specifically, FOR and the combined approach significantly outperformed the BOT and no-training conditions across the three dependent variables; and the combined approach slightly outperformed FOR on the variables of inter-rater reliability and correlation accuracy. The results implied that assessor training is essential to the assessment center process, with the components of FOR being incorporated into the training design, and as long as the assessors are trained with an effective strategy, regardless of the perceived purpose of the assessment (i.e., selection or development) reliable and accurate ratings can be obtained.

摘要

本研究旨在探討評估培訓方法 [assessor training strategy] 與評估目的 [perceived purpose of the assessment] 對總體行爲表現評估準確性 [overall rating accuracy]（即評估師之間的信度 [inter-rater reliability]，關聯精度 [correlation accuracy] 和離差精度 [deviation accuracy]）的影響。評估培訓方法 [assessor training strategy] 分爲四個組別：行爲觀察培訓 [behavioral observation training]，基準參考培訓 [frame-of-reference training]，綜合培訓 [behavioral observation training 和 frame-of-reference training] 及對照組 [no-training]；而評估目的 [perceived purpose of the assessment] 分爲三個組別：人事遴選 [personnel selection]，反饋 [developmental feedback] 及實驗研究 / 對照組 [research-only / control condition]。

二百四十位香港大學生分別隨機分到共十二個實驗組與對照組，並對兩段錄製面談作出評估。是次研究亦邀請四名資深人力資源從業者爲專家小組，其評估會用作爲"標準分" [standard score] 去比較實驗參與者的評估，從而產生實驗參與者的三個總體行爲表現評估準確性 [overall rating accuracy] 的指標。

實驗分析結果顯示只有評估培訓方法 [assessor training strategy]對總體行爲表現評估準確性 [overall rating accuracy] 的影響有達到顯著差異，而主要差異出現在基準參考培訓 [frame-of-reference training] 及綜合培訓 [behavioral observation training 和 frame-of-reference training] 的總體行爲表現評估準確性

[overall rating accuracy] 分別顯著地優於行爲觀察培訓 [behavioral observation training] 和對照組 [no-training]。另外，綜合培訓 [behavioral observation training 和 frame-of-reference training] 在評估師之間的信度 [inter-rater reliability] 及關聯精度 [correlation accuracy]稍爲優於基準參考培訓 [frame-of-reference training]。

研究結果意味著評估培訓 [assessor training] 對評價中心 [assessment centre] 是必不可少的，並且應採用基準參考培訓 [frame-of-reference training] 的要素作爲設計基礎以提升總體行爲表現評估準確性 [overall rating accuracy]。此外，研究結果亦提出無論是基於什麼評估目的 [perceived purpose of the assessment]，只要評估師 [assessor] 是接受有效的評估培訓 [assessor training]，評定出來的行爲表現評估是可以說是可靠及準確的。

Acknowledgements

I wish to thank my examination committee, Dr. Timothy Patrick Carey, Professor Winton Au, Professor Darius Chan, Professor Chan Wai and Professor Norbert Tanzer for their intellectual support and comments on the preparation of this manuscript. Their insightful input and excellent guidance made this production of the thesis more rewarding. I also wish to thank Joyce Kwan, Ray Cheung and Wendy Li for statistical support given. In addition, I would like to express my gratitude to Gina Marescia, Irene Fok, Peng Fu and Natalie Carolan for participating the study as the expert rating panel.

Table of Contents

Table of Contents (Cont'd)

List of Tables

Assessor Training: Influence of Training Strategy and Perceived Purpose of the Assessment on Overall Rating Accuracy

Chapter 1: Introduction

Since the 1940s, assessment center methodology has become a well-accepted technique for personnel selection and development. A meta-analysis conducted by Schmidt and Hunter (1998) indicated that the assessment center methodology possesses moderate predictive validity of .37. Yet, this popular selection procedure has always been criticized on its low construct validity (Achambeau, 1979; Goodstone & Lopez, 2001; Lievens, 1998; Neidig, Martin & Yates, 1979; Sackett & Dreher, 1982; Sackett & Harris, 1988; Schleicher, Day, Mayes & Riggio, 2002; Schneider & Schmitt, 1992). Researchers have suggested numerous interventions that can be employed to promote construct validity, and one of them is carrying out assessor training prior to the assessment center program for enhancing the quality of the ratings (Bray & Grant, 1966; Norton, 1981; Schleicher, Day, Mayes & Riggio, 2002; Silverman, Dalessio, Woods & Johnson, 1986; Spychalski, Quinones, Gaugler, & Pohley, 1997; Thomson, 1970; Turnage & Muchinsky, 1982).

Assessor Training Strategy

Along these lines, the Guidelines and Ethical Considerations for Assessment Center Operations (International Task Force on Assessment Center Guidelines, 1989

& 2000) suggested that assessor training is a requisite for the assessment center program in order to increase validity. Over the past few decades, there are four major strategies that have been used by trainers. These are: (a) rater error training, (b) performance dimension training, (c) behavioral observation training, and (d) frame-of-reference training. Unfortunately, the guidelines do not specify which approach is more effective at improving rating accuracy, stating only that

> "Whatever the approach to assessor training, the objective is to obtain reliable and accurate assessor judgments. A variety of training approaches may be used, as long as it can be demonstrated that reliable and accurate assessor judgments are obtained." (International Task Force on Assessment Center Guidelines, 2000, p. 325).

Thus, one of the goals of the present study is to investigate the effectiveness of assessor training strategy on overall rating accuracy.

In current practice, human resource practitioners and trainers are less likely to focus the entire training sessions solely on familiarizing the assessors with common rater errors (e.g., halo, central tendency, leniency, etc.) including encouraging them to avoid those errors; instead they are inclined to incorporate the rater error training as part of the behavioral observation training strategy, based on the argument that occurrence of rater error is the result of problems taking place during the observation

stage of the assessment process (Thornton & Zorich, 1980). In addition, Ballantyne & Povah (2004) put forward the idea that frame-of-reference training strategy is an extended version of the performance dimension training strategy – instead of just familiarizing the assessors with the performance dimensions, human resource practitioners also train the assessors to be aligned on the evaluation standard that used during the assessment. Consequently, to the extent possible and practical, the efficacies of the behavioral observation training strategy and frame-of-reference training strategy on overall rating accuracy are the major focuses in this study.

Overview of behavioral observation training. According to Pulakos (1986), behavioral observation training is derived from behavior-driven or data-driven theory. Assessors are assumed to be capable of observing specific behaviors, recording them in detail, classifying them into different dimensions and evaluating them based on accurate judgment. This proposition suggests there are distinct stages for the entire assessment process. Similarly, Thornton and Zorich (1980) argued that observation and judgment are two distinct processes, wherein observation is made up of the course of actions like uncovering, identifying and recalling behaviors, whilst judgment comprises the processes of classification, assimilation and evaluation. Along these lines, Byham (1977) introduced a type of behavioral observation training that emphasizes the unique elements in each of the four rating processes (i.e.,

observation, recording, classification and evaluation, ORCE). Assessors being trained under this approach have to closely follow the sequence of these four stages, and can proceed to the next stage only after the previous stage is completed (Lievens, 2001).

Nevertheless, the effectiveness of behavioral observation training has received little research attention in the past few decades (MacDonald & Sulsky, 2009; Woehr & Huffcutt, 1994). As reported in the meta-analysis conducted by Woehr and Huffcutt (1994), there were only four studies evaluating the effectiveness of this behavioral observation training strategy (ORCE). Despite this, the findings of the studies which compared this approach with a no-training condition were quite positive, with the mean effect size of rating accuracy at $d = 0.77$. Hence, the first hypothesis (H1A) is that assessors under the behavioral observation training condition are expected to produce more accurate ratings than assessors under a no-training (control) condition.

Overview of frame-of-reference training. As opposed to behavioral observation training, frame-of-reference training originated from another cognitive-based theory, known as schema-driven theory (Pulakos, 1986). This training strategy does not emphasize the differentiation between observation and evaluation; rather its primary goal is to align assessors' conceptualizations of (a)

competencies or performance dimensions that are being assessed (b) key behavioral indicators along the rating scale within each competency and (c) the evaluation of the behavioral examples and competencies, with the appropriate reference standard suggested by the subject matter experts (e.g., assessment center or development center managers and human resource consultants) (Lievens, 2001; Pulakos, 1984; Sulsky & Day, 1992, 1994). For instance, the assessors being trained with this approach will learn to align with each other in the definition of what communication mean, what behaviors they should look for as regards to communication and what rating should be given if the assessee performed in such a way. In addition to the alignment with a new reference frame, there is a crucial element involved in this approach - discussion and feedback giving. This discussion process among the assessors during the training allows them to clear any doubts they have and collectively work on the same evaluative standard. In this regard, assessors will possess a common collective mental schema regarding the behaviors that are being evaluated, and use that new frame of reference to evaluate the assessees. With this shared conceptualization, the frame-of reference training strategy is suggested to be the most efficacious single training approach for producing reliable and accurate assessment ratings. This is supported by the results found by Woehr and Huffcutt (1994), which indicated that this training methodology has a large mean effect size

on rating accuracy (d = 0.83), as compared to a no-training (control) condition. Thus, in the present study, assessors under a frame-of-reference training condition are expected to produce more accurate ratings than assessors under a no-training (control) condition (Hypothesis 1B).

In addition, Woehr and Huffcutt (1994) in their meta-analysis, found that a frame-of-reference training strategy demonstrated a greater mean effect size (d = 0.83) than a behavioral observation training strategy (d = 0.77) on rating accuracy; hence, assessors under a frame-of-reference training condition are hypothesized to produce more accurate ratings than assessors under a behavioral observation training condition in this study (Hypothesis 2).

Combination training strategies. According to the meta-analysis done by Woehr and Huffcutt (1994), a combination of different assessor training approaches can enhance the utility of assessor training and lead to more accurate assessment results. This was contradicted by a later study (Roch & O'Sullivan, 2003) which suggested that the combined strategies – specifically the approach of incorporating behavioral observation training into frame-of-reference training – did not significantly improve rating accuracy in terms of deviation accuracy and behavioral recall. However, given that human resource practitioners and trainers in the field tend to use combined training strategies when carrying out current assessor training

practices (Ballantyne & Povah, 2004), this study hypothesizes that assessors under a combination training condition (i.e., behavioral observation training and frame-of-reference training) are able to produce more accurate ratings than assessors under a behavioral observation training approach (Hypothesis 3A) or a frame-of-reference training approach (Hypothesis 3B).

In summary, the first series of hypotheses include the following:

H1A: Assessors under a behavioral observation training condition will have a higher rating accuracy across the three dependent variables than assessors under a no-training (control) condition.

H1B: Assessors under a frame-of-reference training condition will have a higher rating accuracy across the three dependent variables than assessors under a no-training (control) condition.

H1C: Assessors under a combined training strategy condition will have a higher rating accuracy across the three dependent variables than assessors under a no-training (control) condition.

H2: Assessors under a frame-of-reference training condition will have a higher rating accuracy across the three dependent variables than assessors under a behavioral observation training condition.

H3A: Assessors under a combined training condition (i.e., behavioral observation training and frame-of-reference training) will have a higher rating accuracy across the three dependent variables than assessors under only a behavioral observation training condition.

H3B: Assessors under a combined training condition (i.e., behavioral observation training and frame-of-reference training) will have a higher rating accuracy across the three dependent variables than assessors under only a frame-of-reference training condition.

Perceived Purpose of the Assessment

In 1971, AT&T adapted the assessment center methodology to hold its first development center (Ballantyne & Povah, 2004). Since then, more and more companies have been using this assessment technique for development purposes (Ballantyne & Povah, 2004). These current human resource practices imply that there are two major purposes for assessment: one is for personnel selection, including both internal promotion and external recruitment; and the other is for development, including the identification of fast track potential, diagnosis of strengths and development areas, feedback giving and succession planning, etc. (Ballantyne & Povah, 2004).

Not surprisingly, the purpose of a given assessment process (i.e., personnel

selection or personal development) may exert an influence on rating accuracy. According to the Process Model of Performance Rating proposed by Landy and Farr (1980) the purpose of an assessment process is a substantial contextual factor impacting both the cognitive rating processes (e.g., observation and judgment) and the organizational rating processes (e.g., internal promotion and career development) which in turn influence the overall rating accuracy and the decision made. A number of studies have examined the effect of perceived purpose of the assessment on overall rating accuracy, yet mixed results have been found (McIntyre, Smith & Hassett, 1984). Some studies have found that ratings used for personnel selection decisions were found to be less accurate than those intended to be used in feedback giving conditions or in conditions where a research purpose was stated up front or even where no information regarding the purpose of the assessment was given (Aleamoni & Hexner, 1980; Bernardin, Orban & Carlyle, 1981; Heron, 1956; Sharon & Barlett, 1969; Taylor & Wherry, 1951; Zedeck & Cascio, 1982). The researchers have speculated that the assessors might go beyond the situation at hand and think of the real-life impact of their ratings (e.g., who should I promote to achieve better team performance without upsetting other potential candidates that working in the same team?) when they make personnel decisions (McIntyre, Smith & Hassett, 1984). As a result, their ratings might be less accurate under this circumstance (i.e., non-research

based assessment centers). However, some researchers have failed to turn up a significant effect for perceived purpose of the assessment on rating accuracy (Berkshire & Highland, 1953; Driscoll & Goodwin, 1979; Gmelch & Glasman, 1977; Hollander, 1957, 1965; McIntyre, Smith & Hassett, 1984). Thus, another goal of the present study is to further investigate the effect of the perceived purpose of the assessment on overall rating accuracy; in particular, the ratings used for personnel selection are hypothesized to be less accurate than those used for giving developmental feedback or for research purposes (Hypothesis 4).

Interaction with Training Strategy

As both assessor training strategy and perceived purpose of the assessment are expected to exert an impact on overall rating accuracy, it is presumed that these two variables may concurrently affect the dependent variables. Yet there are only a few studies that have examined this interaction effect on rating accuracy, and only a slight or no significant interaction effect has been found (McIntyre, Smith & Hassett, 1984; Zedeck & Cascio, 1982). Thus, the final objective of the present study is to further investigate the possible interaction effect of assessor training strategy and perceived purpose of the assessment on rating accuracy (Hypothesis 5).

In summary, the hypotheses regarding perceived purpose of the assessment are as follows:

H4: The ratings used for personnel selection will be less accurate across the three dependent variables than those used for giving developmental feedback or for research purposes (the control condition).

H5: There will be an interaction effect of assessor training strategy and perceived purpose of the assessment on rating accuracy.

Chapter 2: Method

Participants

Two hundred and forty university students (120 males and 120 females) from The Chinese University of Hong Kong were recruited. The reason for recruiting university students as participants was that they likely are not biased regarding the assessor training methodologies examined in this study. As they generally have not been formally trained on assessor skills or performance evaluation skills, and have not evaluated performance in the context of an assessment center or development center program, they are expected to be unbiased. At the same time, they do routinely evaluate professors' and teaching assistants' performance, so they are familiar with rating processes. Thus, because the sample recruited does not possess in-depth knowledge about the assessment training techniques and yet is familiar with a form of rating system, an unbiased training effect can be measured with raters who have some experience.

Experimental Design & Procedures

Participants were randomly assigned to one of the 12 experimental conditions in a four (assessor training strategy) by three (perceived purpose of the assessment) design. There were a total of 20 participants in each condition, (ten males and ten females). After either being trained according to a specific training methodology, or

engaging in a no-training condition, the specific purpose of the assessment was related to the participants. (Note: different purposes of the assessment were instructed with respect to the experimental conditions.) Then they watched and rated the two videotaped interviews. In order to minimize practice effects, the order of videotaped interviews were counterbalanced.

Assessor training strategy. Assessor training strategy was manipulated with four conditions, such that each participant was randomly trained by either behavioral observation training strategy (BOT), frame-of-reference training strategy (FOR), the combination training strategy (BOT & FOR) or assigned to a no-training condition. Before the training session started, participants were informed that they were going to evaluate two interview performances against two performance dimensions (i.e., communication and analysis & problem solving) at the end of the session.

Behavioral observation training (BOT). For this condition, participants were first given an overview of the four distinct sequential processes (i.e., observation, recording, classification and evaluation, ORCE) used to evaluate performance. Subsequently, a copy of the two competency definitions was distributed. The experimenter read aloud the definitions and the associated key behavioral indicators; and the participants were then instructed to observe and record the targeted, specific, competency-related behavioral indicators, as opposed to

making non-behavioral judgments as the first step. They were also taught some essential note-taking skills (e.g., writing shorthand or using symbols) and reminded to avoid some common rater errors (e.g., halo, leniency) to further improve their rating accuracy. Next, a behavioral classification exercise was administered in which the participants categorized 12 behavioral examples (see Appendix 4) into the two performance dimensions to practice the technique of classifying behavioral examples by competency. The experimenter then provided the answers and re-emphasized the importance of recording specific behaviors. Lastly, participants were taught the concept of evaluating each competency with the use of the five-point scale (with 5 being outstanding, 3 being satisfactory and 1 being unsatisfactory) on the provided BARS (Byham, 1977). Participants were then provided with an observation form to record behaviors. This BOT training lasted for approximately 20 minutes. In order to equate the training session time between different training conditions, an overview of the competency-based interview was briefly introduced at the beginning of the experiment.

Frame-of-reference training (FOR). For this condition, a copy of the two-competency definitions was first distributed. The experimenter read aloud the definitions and discussed the importance of forming a common frame-of-reference in assessing performance. Next, the BARS for each of the two competencies were

handed out, and the experimenter discussed the behavioral incidents representing different levels of performance with the participants in order to ensure they understood the concept of how a common frame-of-reference can be developed among assessors using the BARS. Participants then sorted the 12 behavioral examples (see Appendix 5) into the two competencies, and the associated behavioral and performance level to practice the principle of developing a frame-of-reference (note: this is a similar exercise to that used in Woehr, 1994, p. 529). Answers were discussed among participants to clear any doubts they had on the rating standard, and the importance of forming a common frame-of-reference was re-emphasized. Observation forms were provided, but the participants were instructed that it was optional for them to take notes during video watching. This FOR training lasted for approximately 20 minutes. In order to equate the training session time between different training conditions, an overview of the competency-based interview was briefly introduced at the beginning of the experiment.

Combination training (BOT & FOR). The behavioral observation training strategy was integrated with the frame-of-reference training strategy in this condition. Participants were first given an overview of the four distinctive sequential processes (i.e., observation, recording, classification and evaluation, ORCE) used to evaluate performance. Subsequently, a copy of the two-competency definitions was

distributed. The experimenter read aloud the definitions and the associated key behavioral indicators; and participants were then instructed to observe and record the targeted, specific competency-related behavioral indicators, as opposed to making non-behavioral judgments as their first step. They were also taught some essential note-taking skills (e.g., writing shorthand or using symbols) and reminded to avoid some common rater errors (e.g., halo, leniency) to further improve their rating accuracy. The experimenter then discussed the importance of forming a common frame-of-reference in assessing performance and the behaviors representing different levels of performance on the five-point scale with the participants by illustrating the BARS of the two competencies. This helped to ensure the participants understood the principle of classification and the concept of how a common frame-of-reference can be developed among assessors by using BARS. Next, participants sorted the 12 behavioral examples (see Appendix 5) into the targeted two competencies and the associated behavioral and performance levels to practice the principles of developing a frame-of-reference and the classification of behavioral examples (note: this is a similar exercise to that used in Woehr, 1994, p. 529). Answers were discussed among participants to clear any doubts they had on the rating standard, and the importance of forming common frame-of-reference and recording specific behaviors was re-emphasized. Participants were provided with an observation form to record

behaviors. This combination training session lasted for approximately 30 minutes.

No-training. No specific training methodology was used for this condition. However, to equate the overall training session time between different training conditions, the concept of the competency-based interview was introduced. Participants then reviewed the two competency definitions along with the relevant behaviorally anchored rating scales, without any guidance on the observation process or the rating standard. Observation forms were provided, but the participants were instructed that it was optional for them to take notes during video watching. This "no training" condition took about 30 minutes.

Perceived purpose of the assessment. The perceived purpose of the assessment was manipulated by reading one of the following instructions after the training, but before introducing the two videotaped interviews:

Personnel selection instruction. Participants were made to believe that the hypothetical assessees being interviewed were candidates applying for the job of management trainee, and the ratings were used to make personnel selection decisions.

Developmental feedback instruction. Participants were told that the hypothetical assessees being interviewed were fresh graduates currently participating in a one-week interviewing skills workshop, and the ratings were used as feedback

given to those fresh graduates for their future improvement.

Research-purpose instruction. Participants were instructed that the ratings were used for a research study on how individuals evaluate interview performance.

Manipulation Checks

To make certain that the experimental conditions of the assessor training strategy and the perceived purpose of the assessment were effectively manipulated; an eight-item multiple choice questionnaire was designed (see Appendix 6a, b, c & d). The participants filled in the questionnaire at the end of the session, and put it into the collection box anonymously after completion to ensure confidentiality. It was designed to question participants about the training content (e.g., performance dimensions and rating scale) and the instructions given, to ensure that they paid attention to the information conveyed by the experimenter during the training. Only the ratings produced by the participants who responded all eight items correctly were included in the study and processed in the later analysis stage.

Competencies and Behaviorally Anchored Rating Scale

The performance of the assessees was designed to vary along two competencies: communication, and analysis & problem solving (see Appendix 1), which are commonly used to assess graduates across different job functions (Rankin, 2004).

Behaviorally anchored rating scales (BARS) were used to evaluate the

hypothetical assessees' performance as being unsatisfactory, satisfactory or outstanding on a five-point scale on a specific competency as measured in the hypothetical interview (see Appendix 2).

Hypothetical Assessee Performance

Two female assessees were interviewed and videotaped. The same interview questions on the competency of analysis & problem solving were asked across the two assessees (see Appendix 3), while the competency of communication was assessed by the way that the assessees responded to the questions. Each videotaped interview lasted for five minutes.

Expert Panel

Mean expert ratings were used as "standard scores" for the hypothetical assessees' interview performance. The expert panel involved four experienced human resources practitioners who qualify as experts because of their practical experience as assessors. Each had at least ten years' experience in assessing, with a mean of approximately 11 years; moreover, each has a strong academic background that includes familiarity with the literature pertaining to assessment centers. They were provided with the interview questions as well as the definitions and the behaviorally anchored rating scales of the two competencies. All experts independently viewed the two videotaped performances and rated them according to the behaviorally

anchored rating scales.

"Standard scores" were generated by averaging the experts' ratings, which were then compare to the participants' ratings to produce the participants' individual correlation accuracy and deviation accuracy indices.

Independent Variables

The two independent variables were assessor training strategy and perceived purpose of the assessment. Assessor training strategy was manipulated with four conditions: behavioral observation training strategy (BOT), frame-of-reference training strategy (FOR), combination training strategy (BOT & FOR) and no-training condition. The perceived purpose of the assessment variable was measured in three conditions: personnel selection, developmental feedback and a research-only condition.

Dependent Variables

Participants' ratings for two hypothetical assessees (i.e., 12 ratings in total) were used to generate three dependent variables: inter-rater reliability, correlation accuracy and deviation accuracy, as are discussed below.

Inter-rater reliability. Inter-rater reliability among 20 participants across the two hypothetical assessees for each of the 12 experimental conditions was calculated. According to Schmitt (1977) and Schneider and Schmitt (1992), participants in each

condition were treated as items to calculate inter-rater reliability.

Correlation accuracy. Correlation accuracy was calculated by averaging the standardized correlation value between the participant's ratings and the "standard score" (i.e., the means of the experts' ratings) across the two hypothetical assessees. The greater the magnitude of the correlation accuracy, the more accurate the ratings will be.

Deviation accuracy. Deviation accuracy was calculated by averaging the absolute value of the deviation of participant's ratings from the "standard score" (i.e., the means of the experts' ratings) across the two hypothetical assessees. The smaller the magnitude of the deviation accuracy, the more accurate the ratings will be.

Two forms of accuracy were reported in the current study as they provided different types of information. Correlation accuracy measured the parallelism between participants' and experts' ratings; whilst deviation accuracy measured the distance of scores between participants' and experts' ratings (McIntyre, Smith & Hassett, 1984).

Proposed Covariate

Unlike human resource practitioners, university students (i.e., the participants) do not have experience in performance evaluation in the context of assessment centers; yet they still have experience in course evaluation across semesters - in

evaluating the performance of both professors and teaching assistants. This raised the concern of controlling the variable of participants' experiences in performance rating as past studies had suggested that assessors' experience in evaluating performance is a critical, positive factor for obtaining rating accuracy (Borman, 1978; Cardy, Bernardin, Abbott, Senderak & Taylor, 1987; Kozlowski, Kirsch & Chao, 1986; Kozlowshi & Mongilio, 1992; Lievens, 2001). However, since this study did not intend to investigate the effect or assessor's experience on rating accuracy, experience in performance evaluation was proposed to be the covariate and to be controlled under this experimental design.

This non-manipulated covariate was estimated by the total number of courses that the participants have taken in their university study, based on the assumption that the more courses they had gone through, the more opportunities they would have had to complete course evaluations and therefore the more experienced they were in rating performance. The information for the number of courses the participants have studied was collected via the manipulation check questionnaire (see Appendix 6a-d).

Analysis

Inter-rater reliability for each of the 12 experimental conditions was calculated. In addition, a four (assessor training strategy) by three (perceived purpose of the assessment) analysis of covariance (i.e., ANCOVA), with participants' experience in

performance evaluation as the proposed covariate, was planned to analyze the correlation accuracy and deviation accuracy indices as the dependent variables. As such, main effects and interaction effects of assessor training strategy and perceived purpose of assessment on these two dependent variables would then be examined.

Before carrying out the analysis of covariance, assumptions of linearity, homogeneity of regression slopes and reliability of covariate had to be checked. With this sample of 240 university students, linearity and reliability of covariates for the dependent variables of correlation accuracy and deviation accuracy were not assumed (see Table 1).

Table 1

Correlation between the proposed covariate and the two dependent variables

	Assessor Training Strategy							
					Combination (BOT & FOR)			
Perceived purpose of the assessment	BOT		FOR				No-training	
	CA[#]	DA[##]	CA[#]	DA[##]	CA[#]	DA[##]	CA[#]	DA[##]
Personnel Selection	0.320	-0.096	0.147	-0.028	0.300	0.257	0.381	-0.157
Developmental Feedback	0.462	-0.469	0.212	-0.024	0.461	0.170	0.433	-0.361
Research-only	0.347	0.305	0.448	-0.178	0.168	0.164	0.439	0.377

Note. CA[#] = Correlation Accuracy, DA[##] = Deviation Accuracy.

Although homogeneity of regression slopes was assumed for the dependent variables of correlation accuracy and deviation accuracy (see Table 2), there was no significant linear relationship between the total number of courses that the participants have studied in the university (i.e., proposed covariate) and the two

dependent variables. In addition, the proposed covariate was not significantly reliable, as a result, a four (assessor training strategy) by three (perceived purpose of the assessment) analysis of variance (i.e., ANOVA) was carried out to analyze the variables.

Table 2

Homogeneity of regression slopes for the two dependent variables

	df		*F*-value		*p*-value	
	CA[#]	DA[##]	CA[#]	DA[##]	CA[#]	DA[##]
Assessor Training Strategy	3	3	1029.955**	173.794**	<0.001	<0.001
Perceived purpose of the assessment	2	2	1.118	1.945	0.329	0.146
Total number of courses (Proposed covariate)	1	1	19.562	0.091	<0.001	0.763
Assessor Training Strategy x Total number of courses	3	3	1.316	0.283	0.271	0.838
Perceived purpose of the assessment x Total number of courses	2	2	0.147	1.678	0.864	0.190
Assessor Training Strategy x Perceived purpose of the assessment	6	6	0.535	1.506	0.781	0.179
Assessor Training Strategy x Perceived purpose of the assessment x Total number of courses	6	6	0.103	1.119	0.996	0.353

Note. CA[#] = Correlation Accuracy DA[##] = Deviation Accuracy

** $p < 0.001$.

Chapter 3: Results

Effect of Gender of Assessor (Participant)

In addition to ensuring all the experimental conditions were gender-balanced, an independent sample t-test for the two gender groups in each experimental condition for the two dependent variables (i.e., correlation accuracy and deviation accuracy) was carried out to check for the effect of gender of assessor (i.e., participant). This variable did not exert a significant effect on the two dependent variables across the experimental conditions (see Table 3 and Table 4).

Table 3

Effect of Gender of Assessor (Participant) on the Correlation Accuracy

	Assessor Training Strategy							
Perceived purpose of the assessment	BOT		FOR		Combination (BOT & FOR)		No-training	
	t-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value
Personnel Selection	0.284	0.780	-0.532	0.601	0.280	0.783	0.201	0.843
Developmental Feedback	0.244	0.810	0.908	0.376	-0.578	0.570	0.731	0.474
Research-only	0.206	0.839	-1.693	0.108	0.085	0.934	1.572	0.133

Table 4

Effect of Gender of Assessor (Participant) on the Deviation Accuracy

	Assessor Training Strategy							
Perceived purpose of the assessment	BOT		FOR		Combination (BOT & FOR)		No-training	
	t-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value	*t*-value	*p*-value
Personnel Selection	-1.597	0.128	-0.780	0.445	<0.001	1.000	0.430	0.672
Developmental Feedback	0.684	0.503	0.293	0.773	-0.980	0.340	-1.246	0.229
Research-only	<0.001	1.000	-0.548	0.591	<0.001	1.000	-0.146	0.885

Expert Ratings

Mean expert ratings were used as "standard scores" to compare with the participants' ratings and generate the participants' individual correlation accuracy and deviation accuracy indices. Mean and standard deviation of expert ratings across the two competencies on each hypothetical assessee are presented in Table 5.

Table 5

Mean and standard deviation of expert ratings across the two competencies on each hypothetical assessee

	Competency											
	Communication						Analysis & Problem Solving					
	Indicator a		Indicator b		Indicator c		Indicator a		Indicator b		Indicator c	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Assessee 1	2.25	0.50	2.75	0.50	3.25	0.50	1.75	0.50	2.50	0.58	3.00	0.82
Assessee 2	2.50	0.58	2.25	0.50	2.75	0.50	2.75	0.50	2.25	0.50	2.75	0.50

Note. M = Mean expert ratings. SD = Standard deviation of expert ratings.
Ratings were in the range of 1 to 5, with 1 being the lowest score and 5 being the highest score.

The inter-rater reliability among the four experts' ratings for Assessee 1 and Assessee 2 was found to be 0.71 and 0.65 respectively.

Mean Correlation Accuracy and Deviation Accuracy

With the mean expert ratings and participants' ratings, the correlation accuracy and deviation accuracy indices for each participant across the 12 experimental conditions were computed. Mean correlation accuracy and deviation accuracy across experimental conditions are presented in Table 6.

Table 6

Mean correlation accuracy and deviation accuracy

Perceived purpose of the assessment	Assessor Training Strategy							
	BOT		FOR		Combination (BOT & FOR)		No-training	
	CA[#]	DA[##]	CA[#]	DA[##]	CA[#]	DA[##]	CA[#]	DA[##]
Personnel Selection	-0.179	0.490	0.837	0.302	0.898	0.288	-1.538	0.848
Developmental Feedback	-0.243	0.490	0.840	0.298	0.892	0.308	-1.536	0.865
Research-only	-0.146	0.496	0.832	0.300	0.899	0.288	1.558	0.854

Note. CA[#] = Correlation Accuracy, DA[##] = Deviation Accuracy.

The average correlation between the correlation accuracy and deviation accuracy among the 12 experimental conditions was found to be -0.271 with the range of -0.166 to -0.428.

Inter-rater Reliability

Inter-rater reliability across the two hypothetical assessees for each of the 12 experimental conditions was calculated and is reported in Table 7.

Table 7

Inter-rater reliability for the 12 experimental conditions

Perceived purpose of the assessment	Assessor Training Strategy			
	BOT	FOR	Combination (BOT & FOR)	No-training
Personnel Selection	0.403	0.777	0.831	0.120
Developmental Feedback	0.406	0.789	0.821	0.115
Research-only	0.409	0.786	0.827	0.100

The results revealed that there was a major difference in inter-rater reliability across the variable of assessor training strategy, but not for the variable of perceived

purpose of the assessment. In particular, participants' ratings were more reliable if they had been trained, regardless of the assessor training strategy being used. In addition, as compared to the behavioral observation training approach, participants in both the frame-of-reference training and the combination training conditions rated the performance more reliably. However, there was no major difference in inter-rater reliability between the frame-of-reference training and combination training approaches.

Correlation Accuracy

The four (assessor training strategy) by three (perceived purpose of the assessment) analysis of variance was carried out with the dependent variable of correlation accuracy. Results are reported in Table 8.

Table 8

ANOVA results for correlation accuracy

	df	*F*-value	*p*-value
Assessor Training Strategy	3	5187.494**	<0.001
Perceived purpose of the assessment	2	0.542	0.582
Assessor Training Strategy x Perceived purpose of the assessment	6	0.953	0.458

** $p < 0.001$.

The results revealed that there was no significant interaction effect of assessor training strategy and perceived purpose of the assessment on correlation accuracy. In addition, there was no significant main effect of perceived purpose of the assessment on correlation accuracy as well. However, a significant main effect of assessor

training strategy was found, F (3, 228) = 5187.494, p <0.001.

To further investigate the significant main effect of assessor training strategy, a post-hoc test was carried out. The results are reported in Table 9.

Table 9

Post-hoc result for the main effect of assessor training strategy on correlation accuracy

	Assessor Training Strategy (J)					
	FOR		Combination (BOT & FOR)		No-training	
Assessor Training Strategy (I)	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value
BOT	-1.026**	<0.001	-1.086**	<0.001	1.355**	<0.001
FOR	-	-	-0.059	0.052	2.381**	<0.001
Combination (BOT & FOR)	-	-	-	-	2.441**	<0.001

Note. Negative mean difference suggested that "Assessor Training Strategy - I row" had lower correlation accuracy than "Assessor Training Strategy - J column".
** $p < 0.001$. Bonferroni.

The post-hoc results indicated that participants who did not go through assessor training had significantly lower correlation accuracy as compared to others who had been trained. The participants who had been trained under the frame-of-reference approach and the combination training approach (BOT & FOR) produced significantly more accurate ratings than the participants trained only under the behavioral observation training strategy. However, the ratings produced by the participants in the combination training condition were slightly more accurate (with a marginal significance) than the ratings produced by the participants in the

frame-of-reference condition.

Deviation Accuracy

A four (assessor training strategy) by three (perceived purpose of the assessment) analysis of variance was carried out on the dependent variable of deviation accuracy. Results are presented in Table 10.

Table 10
ANOVA results for deviation accuracy

	df	*F*-value	*p*-value
Assessor Training Strategy	3	809.547**	<0.001
Perceived purpose of the assessment	2	0.283	0.754
Assessor Training Strategy x Perceived purpose of the assessment	6	0.208	0.974

** $p < 0.001$.

The results revealed that there was no significant interaction effect of assessor training strategy and perceived purpose of the assessment on deviation accuracy. In addition, there was no significant main effect of perceived purpose of the assessment on deviation accuracy as well. However, a significant main effect of assessor training strategy was found, $F(3, 228) = 809.547, p < 0.001$.

To further investigate the significant main effect of assessor training strategy, a post-hoc test was carried out. The result is reported in Table 11.

Table 11

Post-hoc result for the main effect of assessor training strategy on deviation accuracy

	Assessor Training Strategy (J)					
	FOR		Combination (BOT & FOR)		No-training	
Assessor Training Strategy (I)	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value	Mean Difference (I-J)	p-value
BOT	0.192**	<0.001	0.198**	<0.001	-0.364**	<0.001
FOR	-	-	0.006	1.000	-0.556**	<0.001
Combination (BOT & FOR)	-	-	-	-	-0.561**	<0.001

Note: Positive mean difference suggested that "Assessor Training Strategy - I row" had lower deviation accuracy than "Assessor Training Strategy - J column".
** $p < 0.001$. Bonferroni.

The post-hoc result demonstrated that participants who did not go through assessor training had significantly lower deviation accuracy compared to others who had been trained. The participants who had been trained under the frame-of-reference approach and the combination approach (BOT & FOR) produced significantly more accurate ratings than the participants trained only under the behavioral observation training strategy. However, there was no significant difference between the deviation accuracy of ratings produced by the participants in the combination training condition and the frame-of-reference condition.

Chapter 4: Discussion

In sum, Table 12 comprises the results as related to the hypotheses:

Table 12

Summary of results related to the hypotheses

Hypotheses		Result
1A	Assessors under a behavioral observation training condition will have a higher rating accuracy across the three dependent variables than assessors under a no-training (control) condition.	Statistically Supported
1B	Assessors under a frame-of-reference training condition will have a higher rating accuracy across the three dependent variables than assessors under a no-training (control) condition.	Statistically Supported
1C	Assessors under a combined training strategy condition will have a higher rating accuracy across the three dependent variables than assessors under a no-training (control) condition.	Statistically Supported
2	Assessors under a frame-of-reference training condition will have a higher rating accuracy across the three dependent variables than assessors under a behavioral observation training condition.	Statistically Supported
3A	Assessors under a combined training condition (i.e., behavioral observation training and frame-of-reference training) will have a higher rating accuracy across the three dependent variables than assessors under only a behavioral observation training condition.	Statistically Supported
3B	Assessors under a combined training condition (i.e., behavioral observation training and frame-of-reference training) will have higher rating accuracy across the three dependent variables than assessors under only a frame-of-reference training condition.	Tentatively supported in dependent variables of inter-rater reliability & correlation accuracy
4	The ratings used for personnel selection will be less accurate across the three dependent variables than those used for giving developmental feedback or for research purposes (the control condition).	Statistically not supported

5	There will be an interaction effect of assessor training strategy and perceived purpose of the assessment on rating accuracy.	Statistically not supported

With respect to Hypotheses 4 and 5, neither a significant main effect of perceived purpose of the assessment nor an interaction effect of assessor training strategy and perceived purpose of the assessment on correlation accuracy and deviation accuracy was found. In addition, there was no major difference in inter-rater reliability across the three levels of perceived purpose of the assessment. Thus, from the results, we cannot say that the perceived purpose of personnel selection will lead to a more or less accurate rating.

Whilst the results of inter-rater reliability, correlation accuracy and deviation accuracy provided support for Hypothesis 1A, 1B and 1C such that participants who did not go through assessor training produced significantly less accurate and reliable ratings compared to those who had been trained. These suggest that as long as the assessors are trained before they deliver their performance ratings, (regardless of the perceived purpose of the assessment) more reliable and accurate ratings could be obtained. In other words, assessor training, but not the perceived purpose of the assessment, is an important factor for obtaining accurate and reliable ratings. Such an implication reinforces the importance and the necessity of conducting assessor training before the assessment center or development center program and is consistent with previous research on assessment center training (Lievens, 2001; Task

Force on Assessment Center, 2000).

The next question was which approach should be used to train assessors to better equip them to produce more reliable and accurate ratings. This study contrasted the effectiveness of the behavioral observation training approach to frame-of-reference training and the combination of these two approaches.

With respect to the effectiveness of the behavioral observation training approach, this study showed that the ratings produced by the participants who had been trained with this strategy were more reliable and accurate than the ratings produced by participants who had not been trained at all. However, the results of the inter-rater reliability, correlation accuracy and deviation accuracy supported Hypothesis 2 that the participants who had been trained with the behavioral observation approach underperformed their counterparts in the frame-of-reference conditions in terms of their rating accuracy. Similar to the meta-analytic results found by Woehr and Huffcutt (1994), the current study showed that training assessors with the techniques of observation and the concept of a four-step behavioral evaluation process was quite effective, but was not sufficient to equip them to yield an even more reliable and accurate rating.

In addition, parallel with the meta-analytic findings from Woehr and Huffcutt (1994), the results across the three dependent variables also clearly demonstrated the

effectiveness of the frame-of-reference strategy as a single approach to produce reliable and accurate ratings through training the assessors to form, align with and use the common collective mental schema over the behaviors and the standard being evaluated.

After investigating the effectiveness of a single-method assessor training approach, the final question addressed in this study was whether a combination training approach would add value to the rating accuracy compared to a single-method approach. Past studies have not drawn solid conclusions with respect to the effectiveness of a combined training approach (Roch and O'Sullivan, 2003; Woehr and Huffcutt, 1994).

The results of the three dependent variables provided support for Hypothesis 3A that participants who had been trained with the combined approach (BOT & FOR) produced more reliable and accurate ratings compared to others who received only behavioral observation training. However, there was no concrete evidence for Hypothesis 3B that the combined training approach was better than the most effective single training approach – frame-of-reference training strategy – in terms of rating accuracy. Yet a tentative conclusion could still be drawn in the way that the combination training approach (BOT & FOR) might add value to rating accuracy beyond the frame-of-reference strategy, as the participants in the combined approach

condition showed a slightly greater inter-rater reliability than their counterparts in the frame-of-reference training condition, as well as a marginally significant mean difference of correlation accuracy between these two approaches.

This tentative conclusion seemed to be different from that of Roch's & O'Sullivan's study (2003); they concluded that the approach of incorporating behavioral observation training into frame-of-reference training – did not significantly improve rating accuracy. Nonetheless, one point to note is that the focuses of their study were on deviation accuracy and behavioral recall accuracy; whilst the current study investigated the effect of combined training approach on correlation accuracy, deviation accuracy and inter-rater reliability. Hence, the current study did provide support to Roch's & O'Sullivan's study (2003) in the way that there was no significant result was found with respect to the deviation accuracy; on the other hand, the results of correlation accuracy and inter-reliability also somewhat supported the meta-analytic findings from Woehr and Huffcutt (1994) that the combined training approach can enhance the utility of assessor training and lead to more accurate assessment results.

In other words, to a certain extent, assessors being trained under the combined training approach can produce ratings that are even more accurate than those from the assessors equipped with the frame-of-reference techniques. One possible reason

is that assessors who were trained with the combined approach had a much clearer mindset about what behaviors they had to look for during the observation stage owing to the alignment formed by the frame-of-reference approach. They also stayed more focused and avoided making non-behavioral judgments during the observation and note-taking stages due to both their understanding of the ORCE from the behavioral observation training, and their ability to refer to the observation notes they wrote, classified and rated the performance according to the aligned rating standard. As a result, with the help of the four-step behavioral observation process and the alignment on the behavior and rating standard, an even more accurate rating might have been obtained.

In sum, the current study somewhat demonstrated that training assessors in the alignment on the definitions of the competencies and the evaluation standards, as well as in the techniques of observation and note-taking would increase the rating accuracy compared to either training approach used alone. These results provide some support for human resources practitioners and trainers to try out or continually use the combined approach to train assessors to yield more reliable and accurate assessment ratings.

Limitations and Future Directions

Although the current study provided support for the positive influence of the

assessor training approach on overall rating accuracy, a number of methodological issues should be addressed in future research.

As university students were used as participants for the current study, generalizability of the positive results of the assessor training approach on overall rating accuracy could be a concern. For organizational settings, assessor training and the assessment ratings are the ingredients for the assessment center programs, and many assessors in assessment center programs are internal human resource practitioners, line managers or external human resource consultants. Nevertheless, the participants for the current study were university students, who would not be in the role of assessing candidates' performance in assessment center program for at least another two to three years. This may impact whether the positive effect of assessor training approach could be generalized to the organizational settings. On the other hand, as the purpose of the study was to examine the unbiased training effect on overall rating accuracy, recruiting university students could be an advantage here as they generally have not been formally trained on assessor skills or performance evaluation skills, while still being familiar with the rating task itself. Therefore, because the sample recruited does not possess in-depth knowledge about the assessment training techniques and yet is familiar with a form of rating system, the positive results of assessor training approach on rating accuracy could be considered

unbiased.

Second, the size of the expert panel was relatively small with only four experienced assessors involved. These assessors were invited based on their significant practical experience in assessing performance in the context of assessment center programs. As noted above, they averaged approximately 11 years of experience in this area. In addition, they were familiar with the literature pertaining to assessment centers. Thus, the quality of their ratings as the "standard scores" is likely high, in spite of the small number of assessors involved. However, in general, having more expert raters is likely preferable to the few included here.

Due to time restrictions, all the training sessions were brief (i.e., approximately 30 minutes) which only allowed the core components of each assessor training strategy to be incorporated into the experimental design (e.g., an overview of strategy). While a longer training period is preferable, (some of the expert raters reported doing full-day trainings for assessors before centers) nonetheless, positive results of assessor training approach on rating accuracy were indeed found. In addition, Dugan (1988) showed that neither the length of assessor training nor amount of assessor training seemed to be a crucial factor regarding rating accuracy (Lievens, 1998). This might imply that the assessor training might not need to be lengthy; as long as it provides the crucial elements of the frame-of-reference training

approach or the combined approach (BOT & FOR), rating accuracy and reliability can be expected. Further studies should revisit the relationship between the length of assessor training and the training approach on rating accuracy to verify this finding.

Finally, although experience of participants in evaluating professors' and teaching assistants' performance was proposed to be the covariate for the current study, no significant linear relationship between this proposed covariate with the two dependent variables was found. This went against past findings that assessors' experience in evaluating performance positively influences rating accuracy (Borman, 1978; Cardy, Bernardin, Abbott, Senderak & Taylor, 1987; Kozlowski, Kirsch & Chao, 1986; Kozlowshi & Mongilio, 1992; Lievens, 2001). One possible explanation of this non-significant relationship was that the current participants' experience in performance rating was estimated by the number of courses they have studied. However, nearly 25% of the participants reported incorrectly on this data point causing that data to be classified as missing or invalid, since the participants provided the total number of course credits they had completed, instead of the total number of courses taken. As a result, in future research, clearer instructions should be made in this area to re-examine the effect of the two independent variables on rating accuracy controlling for the participants' experience in performance ratings.

Implications and Conclusion

In summary, the current study investigated the effect of both assessor training approach and the perceived purpose of the assessment on the overall rating accuracy. Along with results of past studies (Lievens, 2001; Task Force on Assessment Center, 2000), the current study re-emphasizes the importance of conducting assessor training before assessment center programs, as this helps assessors to produce more reliable and accurate ratings.

Furthermore, the present study provided positive evidence that both the frame-of-reference training approach and the combination training strategy (BOT & FOR) were superior to a single behavioral observation training approach in producing reliable and accurate ratings. Thus, one of the practical implications of this study is that best practice in assessor training should incorporate the components of the frame-of-reference training strategy. In particular, the assessor training should allow the assessors to align with each other on the behaviors and the standards to be evaluated, as well as encouraging the assessors to use the newly formed common ground to assess performance. In addition, there should be discussion among the assessors which enables them to clear any doubts they have in respect to the evaluative standard.

Moreover, the combination training approach (BOT & FOR) slightly

outperformed the frame-of-reference strategy in inter-rater reliability and correlation accuracy, but not deviation accuracy, these results could still provide justification for human resources practitioners or trainers to train assessors in the four-step behavioral evaluation process as part of the frame-of-reference approach. Specifically, attention should be given to the separation of observation and evaluation, in addition to creating the common frame-of-reference, to yield even more reliable and accurate ratings and in turn to add value to the assessment center program as a whole.

Last but not least, as only the effect of assessor training strategy was found to be significant, this implies that assessors seem not to be affected by the perceived purpose of the assessment. In other words, as long as assessors are trained with an effective strategy (based on this study that would be the frame-of-reference strategy with or without the behavioral observation training approach) before rating assessees' performance, regardless of the perceived purpose of the assessment, reliable and accurate ratings can be obtained.

Appendix 1

Competency Profile

Communication

Definition: Is skilled at creating open communication, with high-impact delivery and effective impression management throughout the process.

Key Behavioral Indicators:

a. Clearly and confidently responds to the questions by giving relevant answers.
b. Delivers messages in a persuasive manner by giving specific and full evidence.
c. Effectively uses appropriate verbal and non-verbal cues to emphasize key facts and/or show enthusiasm and interest.

Analysis & Problem Solving

Definition: Thinks logically and critically about issues and analyses problems in a systematic but timely manner.

Key Behavioral Indicators:

d. Efficiently identifies and analyses the root causes of the targeted issues or problems.
e. Recognizes all important, problem-related information and is able to see the linkages behind the information given.
f. Critically evaluates important information and alternatives, and generates logical and timely solutions.

Appendix 2

Behaviorally Anchored Rating Scales

Communication Is skilled at creating open communication, with high-impact delivery and effective impression management throughout the process.				
1 – Unsatisfactory	**2**	**3 – Satisfactory**	**4**	**5 – Outstanding**
• Gives irrelevant answer.		• Clearly responds to the questions by giving relevant answers.		• Clearly and confidently responds the questions by giving relevant answers that go beyond what was asked.
• Fails to deliver messages in a persuasive manner.		• Occasionally gives specific evidence that is persuasive. Is somewhat persuasive.		• Delivers messages in a very persuasive manner by consistently giving specific and full evidence.
• Shows inappropriate verbal and non-verbal cues (no direct eye contact, foot-tapping, etc.) which appears to be rooted in nervousness and/or boredom and/or arrogance (e.g. long silences, excessive watching of the time).		• Expresses ideas with the use of appropriate verbal and non-verbal cues (e.g. smiles and maintains eye contact)		• Effectively and consistently uses appropriate verbal and non-verbal cues to emphasize key facts and/or show enthusiasm and interest (e.g. maintains smiling and eye contact, and varies voice volume and pitch).

Analysis & Problem Solving Thinks logically and critically about issues and analyses problems in a systematic but timely manner.				
1 – Unsatisfactory	**2**	**3 – Satisfactory**	**4**	**5 – Outstanding**
• Fails to identify any causes of the targeted problems or identifies causes that are unlikely to be accurate.		• Identifies some superficial causes of the targeted problems.		• Identifies the root causes of the targeted problems and fully analyses the problems.
• Fails to seek problem-related information, misses obvious clues when such information is presented.		• Considers different points of view and recognizes some important, problem-related information but misses some clues and/or linkages.		• Considers multiple points of view, recognizes all important, problem-related information; is able to see the linkages behind all sought information.
• Fails to generate logical or relevant solutions when presented with problems.		• Generates mostly useful 1-2 solutions based on the available important information and alternatives.		• Efficiently generates numerous logical and timely solutions based on a complete critical evaluation of all available important information and alternatives.

Appendix 3

Interview Questions – Analysis & Problem Solving

1. Tell me about the most challenging problem that you have recently faced?
   - What was the problem?
   - Why was this challenging?
   - How did you deal with it?
   - What was the outcome?

Appendix 4

Behavioral Classification Exercise – with answer

Tasks: Using the 2 performance dimensions provided (i.e. Communication and Analysis & Problem Solving), indicate the appropriate performance dimension for each example.

Interview Behavioral Example	**Performance Dimensions**
1. Sharon sometimes made eye contact with the interviewer.	Communication
2. Patty revealed that the reason she failed to resolve the customer complaint was partly because she did not recognize the linkage between the production and delivery schedules.	Analysis & Problem Solving
3. David gave examples with specific details regarding the problem and the parties involved.	Communication
4. Tammy would look for the information regarding the budget, product design, and targeted market and production costs before making a decision on launching a new product.	Analysis & Problem Solving
5. When the team was at a 3-3 deadlock in voting, Alex, as a team leader, evaluated the pros and cons of the two ideas and attempted to create a win-win solution.	Analysis & Problem Solving
6. When asked about the most challenging problem she had recently faced, Jess mentioned her sister's problem, instead of her own.	Communication
7. Facing the sudden increase in accident rates, Tom called for an urgent team meeting to understand why this happened and found out that the cause was that some new temporary staff did not strictly follow the safety procedures.	Analysis & Problem Solving
8. Catherine increased her pitch when she emphasized some key facts.	Communication
9. Simon scolded his subordinate and asked how he could have overlooked the targeted customer preference when he planned the commercial media campaign.	Analysis & Problem Solving
10. Ann answered the questions with many fillers (e.g. um…, er….).	Communication
11. Ivy suggested that poor financial performance was partly due to the price being set too high, but failed to recognize that the fundamental problem was that the wrong customer segment was targeted.	Analysis & Problem Solving
12. The interviewer often asked Desmond to repeat his statements, as she said "I don't quite understand what you said".	Communication

Appendix 5

Sorting Exercise – with answer

Tasks: Using the 2 performance dimensions and BARS provided, indicate the appropriate performance dimension, indicator level (item a,b,c) and performance level (1,3,5) for each example.

NOTE: "Key behavioral indicators (item a,b,c)" correspond to the 3 key behavioral indicators indicated in the competency definitions.

"Performance level (1,3,5)" represents "unsatisfactory", "satisfactory" and "outstanding" performance.

Interview Behavioral Examples	**Performance Dimensions**	**Indicator Level** (item a, b, c)	**Performance Level** (1, 3, 5)
1. Sharon sometimes made eye contact with the interviewer.	Communication	c	3
2. Patty revealed that the reason she failed to resolve the customer complaint was partly because she did not recognize the linkage between the production and delivery schedules.	A & P	b	3
3. David gave examples with specific details regarding the problem and the parties involved.	Communication	b	5
4. Tammy would look for the information regarding the budget, product design, and targeted market and production costs before making a decision on launching a new product.	A & P	b	5
5. When the team was at a 3-3 deadlock in voting, Alex, as a team leader, evaluated the pros and cons of the two ideas and attempted to create a win-win solution.	Communication	c	5
6. When asked about the most challenging problem she had recently faced, Jess mentioned her sister's problem, instead of her own.	Communication	a	1

Interview Behavioral Examples	**Performance Dimensions**	**Indicator Level** (item a, b, c)	**Performance Level** (1, 3, 5)
7. Facing the sudden increase in accident rates, Tom called for an urgent team meeting to understand why this happened and found out that the cause was that some new temporary staff did not strictly follow the safety procedures.	A & P	a	5
8. Catherine increased her pitch when she emphasized some key facts.	A & P	c	5
9. Simon scolded his subordinate and asked how he could have overlooked the targeted customer preference when he planned the commercial media campaign.	A & P	b	1
10. Ann answered the questions with many fillers (e.g. um…, er….).	Communication	c	1
11. Ivy suggested that poor financial performance was partly due to the price being set too high, but failed to recognize that the fundamental problem was that the wrong customer segment was targeted.	A & P	a	3
12. The interviewer often asked Desmond to repeat his statements, as she said "I don't quite understand what you said".	Communication	a	1

Appendix 6a

Manipulation Check Questionnaire - BOT

Instruction: Please circle the correct answer.

1. Which 2 competencies / performance dimensions were used to measure the interview performances?
   a. Analysis & Problem Solving + Communication
   b. Communication + Planning Skills
   c. Analysis & Problem Solving + Planning Skills
   d. No specific competency was measured

2. The rating scale we used today was a ____-point scale.
   a. 3
   b. 5
   c. 7
   d. 9

3. According to the instructions, your ratings of the 2 competencies / performance dimensions will be used for_____.
   a. Management Trainee personnel selection
   b. giving developmental feedback to fresh undergraduates
   c. a research study of interview performance
   d. No specific purpose was given

4. The main theme of today's experiment was to evaluate the assessees' interview performances _____.
   a. With the usage of a 4-stage behavioral, sequential process
   b. By forming common ground with other group mates among the definitions and the behavior examples representing different rating levels of performance on each of the 2 competencies
   c. both a & b
   d. No specific theme was illustrated, except to understand the concept of a behavioral interview

5. What is the sequence for the 4-stage assessment process?
   a. Evaluate, Observe, Classify, Record
   b. Classify, Evaluate, Observe, Record
   c. Observe, Record, Classify, Evaluate
   d. Record, Evaluate, Observe, Classify

6. What stages can be suggested to be judgmental?
   a. Evaluate
   b. Record
   c. Classify
   d. Both a & c

7. What of the following is not a common rater error?
   a. Halo
   b. Leniency
   c. Stereotyping
   d. Recording slowly

8. Our goal is to record _____ evidence.
   a. specific
   b. detailed
   c. non-judgmental
   d. All of the above

Background Information:

Gender: M / F

Study year: _____

Total no. of courses you have completed (*excluding the current semester*):_____

Appendix 6b

Manipulation Check Questionnaire - FOR

Instruction: Please circle the correct answer.

1. Which 2 competencies / performance dimensions were used to measure the interview performances?
   a. Analysis & Problem Solving + Communication
   b. Communication + Planning Skills
   c. Analysis & Problem Solving + Planning Skills
   d. No specific competency was measured

2. The rating scale we used today was a ____-point scale.
   a. 3
   b. 5
   c. 7
   d. 9

3. According to the instructions, your ratings of the 2 competencies / performance dimensions will be used for_____.
   a. Management Trainee personnel selection
   b. giving developmental feedback to fresh undergraduates
   c. a research study of interview performance
   d. No specific purpose was given

4. The main theme of today's experiment was to evaluate the assessees' interview performances by_____.
   a. With the usage of a 4-stage behavioral, sequential process
   b. By forming common ground with other group mates among the definitions and the behavior examples representing different rating levels of performance on each of the 2 competencies
   c. both a & b
   d. No specific theme was illustrated, except to understand the concept of a behavioral interview

5. Why we need to form a common frame of reference?
   a. Different people have different subjective interpretations
   b. We need to have a common standard for accurate evaluation
   c. Both a & b
   d. It is not necessary to form a common frame of reference

6. What type of information helps individuals to develop common ground for performance evaluation?
   a. Performance dimension definitions
   b. Behaviors associated with that performance dimension
   c. Behaviors associated with the rating scale
   d. All of the above

7. *"Interviewee 1 suggested that he tried to resolve the client's complaint by gathering information from her perspective, his boss perspective and the organization's refunding policy"* Thus, based on what we have discussed, what competency you would classify this example in?
   a. Communication
   b. Analysis & Problem Solving
   c. All of the above
   d. None of the above

8. With this process, we are trying to compare each individual's performance against _____.
   a. The previous candidate's performance
   b. An evaluation standard
   c. All of the above
   d. None of the above

Background Information:

Gender: M / F

Study year:_____

Total no. of courses you have completed (*excluding the current semester*):_____

Appendix 6c

Manipulation Check Questionnaire – Combined Approach

Instruction: Please circle the correct answer.

1. Which 2 competencies / performance dimensions were used to measure the interview performances?
   a. Analysis & Problem Solving + Communication
   b. Communication + Planning Skills
   c. Analysis & Problem Solving + Planning Skills
   d. No specific competency was measured

2. The rating scale we used today was a ____-point scale.
   a. 3
   b. 5
   c. 7
   d. 9

3. According to the instructions, your ratings of the 2 competencies / performance dimensions will be used for_____.
   a. Management Trainee personnel selection
   b. giving developmental feedback to fresh undergraduates
   c. a research study of interview performance
   d. No specific purpose was given

4. The main theme of today's experiment was to evaluate the assessees' interview performances _____.
   a. With the usage of a 4-stage behavioral, sequential process
   b. By forming common ground with other group mates among the definitions and the behavior examples representing different rating levels of performance on each of the 2 competencies
   c. both a & b
   d. No specific theme was illustrated, except to understand the concept of a behavioral interview

5. What is the sequence for the 4-stage assessment process?
   a. Evaluate, Observe, Classify, Record
   b. Classify, Evaluate, Observe, Record
   c. Observe, Record, Classify, Evaluate
   d. Record, Evaluate, Observe, Classify

6. Why we need to form a common frame of reference?
   a. Different people have different subjective interpretations
   b. We need to have a common standard for accurate evaluation
   c. Both a & b
   d. It is not necessary to form a common frame of reference

7. What type of information helps individuals to develop common ground for performance evaluation?
   a. Performance dimension definitions
   b. Behaviors associated with that performance dimension
   c. Behaviors associated with the rating scale
   d. All of the above

8. What of the following is not a common rater error?
   a. Halo
   b. Leniency
   c. Stereotyping
   d. Recording slowly

Background Information:

Gender: M / F

Study year:_____

Total no. of courses you have completed (*excluding the current semester*):_____

Appendix 6d

Manipulation Check Questionnaire – Control / No-training

Instruction: Please circle the correct answer.

1. Which 2 competencies / performance dimensions were used to measure the interview performances?
   a. Analysis & Problem Solving + Communication
   b. Communication + Planning Skills
   c. Analysis & Problem Solving + Planning Skills
   d. No specific competency was measured

2. The rating scale we used today was a ____-point scale.
   a. 3
   b. 5
   c. 7
   d. 9

3. According to the instructions, your ratings of the 2 competencies / performance dimensions will be used for_____.
   a. Management Trainee personnel selection
   b. giving developmental feedback to fresh undergraduates
   c. a research study of interview performance
   d. No specific purpose was given

4. The main theme of today's experiment was to evaluate the assessees' interview performances _____.
   a. With the usage of a 4-stage behavioral, sequential process
   b. By forming common ground with other group mates among the definitions and the behavior examples representing different rating levels of performance on each of the 2 competencies
   c. both a & b
   d. No specific theme was illustrated, except to understand the concept of a behavioral interview

5. What of the following is an important element of competency based interviews?
   a. Job-relatedness of the interview
   b. Standardization of the process
   c. Structured use of the data to evaluate the candidate
   d. All of the above

6. "Can you tell me about a time when you led a team?" is an example of a(n) _____question.
   a. open-ended
   b. experience-based
   c. situational
   d. None of the above

7. What specific details do CBI interviewers look for when asking about an experience?
   a. The Circumstance
   b. The Impact
   c. Both a & b
   d. None of the above

8. What of the following are the tips you have to beware of to do well in an interview?
   a. Conduct research on the industry that the company is in
   b. Arrive at least 30 minutes early
   c. Make negative comments about the company you worked in previously
   d. None of the above

Background Information:

Gender: M / F

Study year: _____

Total no. of courses you have completed (*excluding the current semester*):_____

References

Aleamoni, L.M. & Hexner, P.Z. (1980). A review of the research on student evaluations and a report on the effects of different sets of instructions on student course and instructor evaluation, *Instructional Science, 9,* 67-84.

Archambeau, D.J. (1979). Relationships among skill ratings assigned in an assessment center, *Journal of Assessment Center Technology, 2,* 7-20.

Ballantyne, I. & Povah, N. (2004). *Assessment and Development Centers* (2nd ed.). England: Gower.

Berkshire, J.R. & Highland, R.W. (1953). Forced choice performance rating: A methodological study, *Personnel Psychology, 6,* 356-378.

Bernardin, H.J., Orban, J.A., & Carlyle, J.J. (1981). *Performance rating as a function of trust in appraisal and rater individual differences* (pp. 311-315). Proceedings of the 41st annual meeting of the Academy of Management.

Borman, W. C. (1978). Exploring the upper limits of reliability and validity in job performance ratings, *Journal of Applied Psychology. 63*, 135-144.

Bray, DW. & Grant, D.L. (1966). The assessment center in the measurement of potential for business management, Psychological Monographs: General & Applied, 80 (17, whole no. 625).

Byham, W.C. (1977). Assessor selection and training. In J.L. Moses & W.C. Byham (Eds.), *Applying the assessment center method* (pp.89-125). New York: Pergamon Press.

Cardy, R.L., Bernardin, H. J., Abbott, J.G., Senderak, M.P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy, *Journal of Occupational and Organizational Psychology, 60,* 197-205.

Driscoll, L.A. & Goodwin, W.L. (1979). The effects of varying information about use and disposition of results on university students' evaluations of faculty and courses, *American Educational Research Journal, 16,* 25-37.

Dugan, B. (1988). Effects of assessor training on information use, *Journal of Applied Psychology, 73, 4,* 743-748.

Gebelein, S.H., Lee, D.G., Nelson-Neuhaus, K.J. & Sloan, E.B (1999). *Successful executive's handbook: Development suggestions for today's executives.* (2nd ed.) Minnesota: Personnel Decisions International.

Goodstone, M.S. & Lopez, F.E. (2001). The frame of reference approach as a solution to an assessment center dilemma, *Consulting Psychology Journal: Practice and Research, 53, 2,* 96-107.

Gmelch, W.H. & Glasman, N.S. (1977). The effects of purpose on student evaluation of college instructors, *Educational Research Quarterly, 2,* 45-55.

Heron, A. (1956). The effects of real-life motivation on questionnaire response. *Journal of Applied Psychology, 40,* 65-68.

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology, 41,* 85-90.

Hollander, E. P. (1965). Validity of peer nominations in predicting a distant performance criterion. *Journal of Applied Psychology, 49,* 434-438.

International Task Force on Assessment Center Guidelines. (2000). Guidelines and ethical considerations for assessment center operations, *Public Personnel Management, 29, 3,* 315-331.

Kozlowski, S.W.J., Kirsch, M.P., & Chao, G.T. (1986). Job knowledge, rate familiarity, conceptual similarity and halo error: A exploration, *Journal of Applied Psychology, 71,* 45-49.

Kozlowski, S.W.J. & Mongillo, M. (1992). The nature of conceptual similarity schemata: Examination of some basic assumptions, P*ersonality and Social Psychology Bulletin, 18,* 88-95.

Landy, F.J. & Farr, J. L. (1980). Performance rating, *Psychological Bulletin, 87, 1,* 72-107.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review, *International Journal of Selection and Assessment, 6,* 141-152.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity, *Journal of Applied Psychology, 86,* 2, 255-264.

Lucia, A.D. & Lepsinger, R. (1999). *The art and science of competency models: Pinpointing critical success factors in organizations*. San Fransisco: Jossey-Bass/Pfeiffer.

MacDonald, H.A. & Sulsky, L.M. (2009). Rating formats and rater training reduz: A context-specific approach for enhancing the effectiveness of performance management, *Canadian Journal of Behavioural Science, 41, 4,* 227-240.

McIntyre, R.M., Smith, D.E & Hassett, C.E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating, *Journal of Applied Psychology, 69, 1,* 147-156.

Neidig, R.D., Martin, J.C. & Yates, R.E. (1979). The contribution of exercise skill ratings to final assessment center evaluations, *Journal of Assessment Center Technology, 2,* 21-23.

Norton, S.D. (1981). The assessment center process and content validity: A reply to Dreher and Sackett, *Academy of Management Review, 6,* 561-566.

Pulakos, E.D. (1984). A comparison of training programs: Error training and accuracy training, *Journal of Applied Psychology, 69,* 182-186.

Pulakos, E.D. (1986). The development of training programs to increase accuracy with different rating tasks, *Organizational Behavior and Human Decision Processes, 38,* 78-91.

Rankin, N. (2004). Degree of attraction: employers' use of competencies in graduate recruitment, *Competency & Emotional Intelligence, 11, 4,* 28-39.

Roch, S.G. & O'Sullivan, B.J. (2003). Frame of reference rater training issues: recall, time and behavior observation training, *International Journal of Training and Development, 7, 2,* 93-107.

Sackett, P.R. & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical finding, *Journal of Applied Psychology, 67,* 401-410.

Sackett, P.R. & Harris, M.M. (1988). A further examination of the constructs underlying assessment center ratings, *Journal of Business and Psychology, 3,* 214-229.

Schleicher, D.J., Day, D.V., Mayes, B.T. & Riggio, R.E (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers, *Journal of Applied Psychology, 87,4,* 735-746.

Schmidt, N. (1977). Interrater agreement in dimensionality and combination of assessment centre judgments. *Journal of Applied Psychology, 62,* 171-176

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings, *Psychological Bulletin, 124(2)*, 262-274

Schneider, J.R. & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs, *Journal of Applied Psychology, 77*, 32-41.

Sharon, A.T. & Bartlett, C.J. (1969). Effects of instructional conditions in producing leniency on two types of rating scales, *Personnel Psychology, 22,* 251-263.

Silverman, W.H., Dalessio, A., Woods, S.B. & Johnson, R.L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39,* 565-578.

Spychalskil, A.C., Quinones, M.A., Gaugler, B.B. & Pohley, K. (1997). A survey of assessment center practices in organizations in the United States, *Personnel Psychology, 50,* 71-90.

Sulsky, L.M. & Day, D.V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77*, 501-510.

Sulsky, L.M. & Day, D.V. (1994). Effect of frame-of-reference training on rater accuracy under alterative time delays, *Journal of Applied Psychology, 79,* 535-543.

Task Force on Assessment Center Guidelines. (1989). Guidelines and ethical considerations for assessment center operations, *Public Personnel Management, 18,* 457-470.

Taylor, E.K. & Wherry, R.J. (1951). A Study of leniency in two rating systems, *Personnel Psychology, 4,* 39-47.

Thomson, H.A. (1970). Comparison of predictor and criterion judgments of managerial performance using the multitrait-multimethod approach, *Journal of Applied Psychology, 54,* 496-502.

Thornton, G.C. & Zorich, R.S. (1980). Training to improve observer accuracy, *Journal of Applied Psychology, 65*, 351-354.

Turnage, J.J. & Muchinsky, P.M. (1982). Transsituational variability in human performance within assessment centers, *Organizational Behavior and Human Performance, 30,* 174-200.

Woehr, D.J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information, *Journal of Applied Psychology, 79*, 525-534.

Woehr, D.J. & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review, *Journal of Occupational and Organizational Psychology, 67,* 189-205.

Zederck, S. & Cascio, W. (1982). Performance decision as a function of purpose of rating and training, *Journal of Applied Psychology, 67,* 752-758.